

**CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL  
DATA: A MACHINE LEARNING APPROACH**

Ryan Henry Papetti

---

A thesis submitted to The Honors College  
In partial fulfillment of the Bachelors of Science degree in  
Information Science and Technology

Fall 2019  
University of Arizona

Approved by: \_\_\_\_\_  
Dr. Richard H. Thompson  
School of Information

## **Abstract**

As the legal cannabis industry emerges from its nascent stages, there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. While many cannabis operators view the state-mandated traceability as a necessary burden, it provides a goldmine for internal customer analysis. Traditionally, segmentation analysis focuses on demographic or RFM (recency-frequency-monetary) segmentation. Yet, neither of these methods has the capacity to provide insight into a customer's purchasing behavior. With the help of 4Front Ventures, a battle-tested multinational cannabis operator, this report focuses on segmenting customers using cannabis-specific data (such as flower and concentrate consumption) and machine learning methods (K-Means and Agglomerative Hierarchical Clustering) to generate newfound ways to explore a dispensary's consumer base. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Although the results are meaningful, this report could benefit with exploring more clustering algorithms, comparing results across dispensaries within the same state, or investigating segmentations in other state markets.

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	The Business Problem . . . . .	4
1.2	Acquisition of Data . . . . .	5
1.3	Scope of Analysis . . . . .	8
<b>2</b>	<b>Customer Segmentation Analysis</b>	<b>9</b>
2.1	Brief Introduction . . . . .	9
2.2	Challenges of Performing Analysis . . . . .	12
<b>3</b>	<b>Clustering Using Machine Learning Methods</b>	<b>14</b>
3.1	Similarity Measures . . . . .	15
3.2	Centroid-based: K-Means . . . . .	16
3.3	Hierarchical-based: Agglomerative . . . . .	23
<b>4</b>	<b>Preparing the Data</b>	<b>29</b>
4.1	Feature Engineering . . . . .	29
4.2	Criteria for Clustering . . . . .	31
4.3	Scaling and Reformatting Data . . . . .	32
<b>5</b>	<b>Performing Analysis and Results</b>	<b>34</b>
5.1	Brief Overview of Code . . . . .	34
5.2	Clustering Results . . . . .	35
5.2.1	K-Means . . . . .	35
5.2.2	Agglomerative . . . . .	38
5.3	Managerial Implications of Results . . . . .	42
<b>6</b>	<b>Future Work and Conclusion</b>	<b>44</b>
6.1	Possible Research Avenues or Expansions . . . . .	44
6.2	Conclusion . . . . .	46

## List of Tables

1	Results of K-Means Clustering with $k = 5$ . . . . .	37
2	Results of Agglomerative Clustering with 6 Clusters . . . . .	40

## List of Figures

1	Random Raw Data . . . . .	17
2	Raw Data with Centroids . . . . .	19
3	Distance from a Random Point to Each Centroid . . . . .	20
4	Update of Centroids after One Iteration . . . . .	21
5	Converged k-means Algorithm . . . . .	22
6	Random Raw Data for Agglomerative Clustering . . . . .	24
7	Sample Distance Calculation from One Point to Each Cluster . . . . .	25
8	Example of a Data Update in Agglomerative Clustering . . . . .	26
9	Dendrogram of Clustered Random Raw Data . . . . .	27
10	Inertia for Several Ks on 4Front (dispensary name redacted) Dispensary Data. . . . .	36
11	Dendrogram for 4Front Dispensary Data . . . . .	39

# 1 Introduction

## 1.1 The Business Problem

Any company in retail, no matter the industry, ends up collecting, creating, and manipulating <sup>1</sup> data over the course of their lifespan. These data are produced and recorded in a variety of contexts, most notably in the form of shipments, tickets, employee logs, and digital interactions. Each of these instances of data describes a small piece of how the company operates, for better or for worse. The more access to data that one has, the better the picture that the data can delineate. With a clear picture made from data, details previously unseen begin to emerge that spur new insights and innovations.

The sheer size and complicated nature of data in the real world make the above task much easier said than done, though. The rise of performance metrics and interactive dashboards have ushered in a new era of looking at data. Many times, the data included in dashboards are at the superficial level: *How much did store X make during December?*, *What are our top 5 products?*, *What is our monthly COGS (Cost of Goods Sold)?*. While dashboards supply data that often have important significance in supply chain management and operations, they are limited in the sense that they omit data and insights that require higher level of data mining and analysis.

Companies that utilize proper data science and data mining practices allow themselves to dig further into their own operating strategies, which in turn allows them to optimize their commercial practices. As a result, there are increasing motivations for investigating phenomena and data that cannot be simply answered: *Why is product B purchased more on the first Saturday of every month compared to other weekends?*, *If a customer bought product B, will they like product C?*, *What are the defining traits of our customers?* *Can we predict what customers will want to buy?* It is the latter half of the last question that will be the broad focus of this paper.

In particular, this paper discusses the results of a customer segmentation analysis project done in conjunction with 4Front Ventures. 4Front Ventures (hereby referred to as 4Front) is a consulting and management firm in the legal cannabis industry that operates various cultivation, production, and retail sites across the country. As 4Front continues expanding into new markets, it is crucial for them to have a sense of who their customers are. Not just the products they like to purchase, but when they like to purchase them, how often they want to purchase them, and what their lifetime value may be to the

---

<sup>1</sup>In general, data manipulation is not malicious nor contains malintent in its nature. It is the simple process of converting data from one format into a more usable, useful one.

company. While some of these questions are more straight forward than others, it is clear that they all require data munging, analysis, and presentation that involve skills and techniques beyond what is required of a traditional analyst.

By integrating machine learning<sup>2</sup> practices and conventional business understandings, the paths to answering these questions became more intertwined with that of a similar question: *What segments or groups of customers do we have?* After studying clustering and reading about it in numerous other contexts, it became clear that clustering 4Front’s retail customers became one way to investigate the purchasing patterns and behaviors of its customers.

## 1.2 Acquisition of Data

Finding readied, usable data for analysis in a business context is a rarity. As such, it is imperative to collect as much data as possible, but also in a format that meets a wide variety of financial, ethical, and computational considerations. But before discussing these, it is first important to describe the ways in which the relevant retail data are stored and utilized across the company.

Without delving into confidential details, the broad idea is that the vast majority of retail data are stored in various SQL databases. Because of emphasis on seed-to-sale traceability, various state regulations, and lack of competition in the software market, most businesses are required to integrate their entire business up to one point of sale (POS) system that is consistent across the company. If the company is vertically integrated, the POS extends to their cultivation and production software. Some software providers, such as Bio-Track, Greenbits, Viridian, have flourished in the industry by providing fully integrated software known as seed-to-sale systems. In the backend, servers store their data in SQL databases built to comply with state regulations and standards. On the front end, they deliver relevant data or insight via interactive dashboards, reporting modules, or simple visuals to retail managers or analysts.

As a direct consequence of 4Front’s successful expansion into new and developing markets, they have incurred unforeseen challenges with data handling and storage. Even though the tactics and strategies that 4Front uses to sell and market their products are, for the most part, consistent across state markets, their data storage and data accessibility is contingent upon their markets and access to third-party software. Certain software, while allowing for nice

---

<sup>2</sup>Generally, machine learning is the science of creating and using models and algorithms that predict or group data in a statistically meaningful way. It is a subset of artificial intelligence (AI).

reports and key visuals, do not have any built-in backend functionality for retailers to access the raw data. Luckily, one of the software used in a couple of 4Front’s operating markets allows for a backend SQL editor that allows for direct queries, though there is very limited documentation on the database structure is provided by the software company. Nonetheless, it is possible for customer data to be collected for customers who exist in the state markets with the appropriate seed-to-sale software.

However, just having access to the data/knowing where it is is a small step in the overall data gathering process. Roughly speaking, it is possible to classify the various data acquisition processes into three distinct categories.

First, it was necessary to establish any ethical considerations or constraints to the usage of data. When first-time customers enter a dispensary, they are presented with a form that asks for verifiable demographic information such as their name, age, and address. In addition, they are also asked if they consent to the company using their data for analysis and marketing purposes. Each customer’s answer to the previous question is one-hot encoded into the database: 0 for “no” and 1 for “yes”. The customers included in this analysis, thus, are only the customers who answered “yes” to the question and have a 1 for the value for the appropriate feature. Furthermore, to protect the anonymity of each of the customers, it is also necessary to prune away all sensitive information from each customer. In other words, the only demographic/sensitive information of each customer that the analysis will use is the age of the customer. The sex, address, name, and other sensitive or personal information is detached from the customer during analysis. Each customer is uniquely identified with an ID that allows for consistent analysis, but the IDs are generated internally, which means that the customer has no knowledge of their ID. Essentially, while there is a way for the program to keep track of a particular customer’s purchases, it is not possible for the program to include customers who do not consent to using their data for this purpose, or for the program to tie the purchases to a particular name or address<sup>3</sup>.

Second, collecting the data in an efficient manner heavily relies on a strong understanding of the structure of the database. Without revealing too many details, there were four important datatables in the database that contained relevant information.

- The *customers* table includes the customer id, number of visits, total amount spent, whether or not they consent to us using their data, and

---

<sup>3</sup>It is possible, though, to tie this information outside of the context of the program. Namely, by accessing the database in a different way without regard to the above considerations.

age of each customer. These data are needed for identifying unique customers and also providing the beginnings of some of the data used in clustering.

- The *tickets* table contained all information regarding tickets <sup>4</sup>, such as the ticket ID, time of transaction, total amount spent, the customer ID involved in the ticket, and which employee completed the ticket. The time of transaction, total amount spent, and associated customer ID are relevant for this particular analysis.
- The *sales* table hosts data related to each individual sale (i.e each individual product sold). This consists of a sales ID, the ticket ID that the sale is associated to, the price of the sale (price of the item), and the product ID associated with the sale. This table contains many IDs and other data that intersect with other tables that are important for this analysis. From this table, it is possible to gather almost all the relevant data for each ticket/customer.
- The *products* table includes the necessary information about each product the store has, such as its ID, when it was added to the system, and which product category <sup>5</sup> it belongs to. This table is mostly used for debugging purposes and for providing some context that makes it easier to identify and classify products.

Lastly, there were certain computational considerations to take into account when collecting data as well. Though the database is set up to handle missing values already, there were several columns in several tables that had malformed or missing values that required additional attention. Incorrect self-reported dates, voided tickets, and tickets with \$0 in sales needed to be pruned from the dataset. In addition, any relevant field with a missing or negative value needed to be pruned or corrected from the dataset. Though the number of affected instances is small, it was crucial to handle these malformed instances because they prevented smooth analysis later on.

After taking the above processes and considerations into account, it was possible to collect the relevant data in a single query using the software's SQL editor. The data was then outputted into a CSV file (with around 250,000 rows) for easy viewing, importing, and analysis.

---

<sup>4</sup>In retail jargon, a ticket is basically a receipt. It is a proof of transaction.

<sup>5</sup>It will be revealed later that the initial product category assignments are incomplete/difficult to parse. It is necessary to collect these now to come up with a smarter way to classify products in the latter parts of the project.



### 1.3 Scope of Analysis

In general, the methods used to gather the data for this project can easily be extended into other relevant contexts/analyses. While there is clear value in using the same data to investigate purchasing patterns or to build an item-based collaborative filtering recommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four intertwined goals:

1. To cluster customers based on common purchasing behaviors for future operations/marketing projects
2. To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines
3. To investigate how similar data and algorithms could be used in future data mining projects
4. To create an understanding and inspiration of how data science can be used to solve real-world problems

Before delving into the details of the project and its implications, the next chapter discusses what customer segmentation analysis actually is and the reasons for its importance.

## 2 Customer Segmentation Analysis

### 2.1 Brief Introduction

For a retailer, understanding the components of their consumer base is key to maximizing their potential in a market; the retailer that attracts the most customers will acquire the most market share, *ceterus paribus*. In fact, the high costs of gaining a new customer or getting back an old customer force retailers to seriously consider how to allocate resources to optimize not just volume of customers, but the retention of them as well<sup>6 7</sup>. Additionally, it is a common understanding in the retail industry that the Pareto Principle—more likely than not—applies to the company: 80% of profits come from 20% of the customers<sup>8</sup>. One crucial reason why this principle holds is because retail businesses thrive on repeat purchases<sup>9</sup>. As a consequence, a net change of one customer can significantly impact a business’ profit in the long run. Therefore, it is generally in the best interest of the retailer to devote efforts to retaining customers by understanding them on as deep of a level as necessary.

However, examining the intricate, rich relationships between a retailer and their consumer base involves understanding how different components of the base behave. Namely, how different segments of customers act similarly or differently from other segments<sup>10</sup>. One method of approaching customer understanding is through the lens of customer segmentation. In short, customer segmentation analysis is the process of grouping customers in such a way that customers within one particular group are similar to each other but different from customers in other groups. In general, there are two paths of segmentation: *a priori* and *post hoc*. *A priori* analysis involves creating the segments beforehand and then, after examining data, placing each customer within the segments<sup>11</sup>. Rather than having the customer data dictate the types of segments formed, certain outside knowledge or structure would dictate the preferred segmentations. As such, the key unit of analysis here are the created segments, not necessarily the customers themselves.

On the other hand, *post hoc* analysis leverages the data to form the segments, rather than the other way around. In a sense, *post hoc* analysis is a direct consequence of advancements in data collection and reliability whereas a

---

<sup>6</sup>Marcus (1998, p. 501)

<sup>7</sup>Cooil et al. (2008)

<sup>8</sup>Zhang (2007)

<sup>9</sup>Marcus (1998, p. 494)

<sup>10</sup>Chen et al. (2012)

<sup>11</sup>Cooil et al. (2008)

prior analysis arose to prominence several years before such beneficial advancements. Regardless of the context, advancing technology has opened doors for post hoc analysis to succeed as a segmentation method in the retail industry. So, modern retailers and data scientists tend to perform customer segmentation using techniques residing under the post hoc umbrella, which will be the focus of the remainder of the paper.

While the goal of customer segmentation analysis has been consistent among retailers for many years, approaches in the past relied on much weaker analytical techniques than available today. It is nonsensical to blame companies in the past who failed to utilize their data properly; the technology and data infrastructure simply were not ubiquitous or cheap enough to allow for companies to collect massive amounts of data as they do today. Yet, many companies still found rudimentary methods to attempt to understand their customers, the most traditional involving purely demographic analysis<sup>12</sup><sup>13</sup>. Demographic analysis is segmenting customers solely on demographic features, such as age, sex, race, or income. It is built upon the assumption that retail behavior is defined by the demographics of the surrounding neighborhood of a store's consumer base. The distillation of customers to only a few well-understood and categorized demographic features meant it was easier for retailers to collect and utilize data from their customers, since it was relatively easy to take a limited number of specific characteristics and generate reasonable predefined categories. Furthermore, demographic analysis also thrived because it became a quick, cheap, and easy model to predict how new customers would interact. So, demographic segmentation allowed for retailers to collect only relevant data—which in turn requires minimal labor and thus cost—that kept analysis and communication of the analysis at a common level. Despite the success of many popular marketing firms, the increasing accessibility of retail technology revealed that demographic segmentation had no capacity to produce insight with consumer purchase histories.

Once retailers and marketing researchers began to tinker with different methods of segmentation, it became clearer sooner rather than later that deeper behavioral segmentation would quickly supersede purely demographic segmentation<sup>14</sup>. Instead of attempting to divide customers based on their demographics, retailers began segmenting their customers based on their purchasing patterns, mostly using a technique known as the Recency-Frequency-Monetary (RFM) method<sup>15</sup>. A standard implementation of the RFM model

---

<sup>12</sup>Marcus (1998, p. 494)

<sup>13</sup>Bhatnagar (2004, p.758)

<sup>14</sup>Bhatnagar (2004, p.758)

<sup>15</sup>Marcus (1998, p. 494)

is cheap and simple: once each of the components are defined in a way that makes them easy to collect, it is a relatively menial task for a retailer to visualize the results, which makes interpretation easy as well. Usually, the results of an RFM analysis would include three plots—one for each combination of two variables (e.g. Recency and Frequency)—with the inferred segments and their defining characteristics. RFM analysis became a staple of modern marketing for its simplicity and its cheap cost to implement as well to communicate efficiently<sup>16</sup>. In a way, the visualization aspect alone gave utility to the RFM model, allowing managers to effectively glean insights from the analysis.

Yet, as the retail industry evolved in parallel with the technology boom, it became dramatically easier for retailers to collect data at a larger scale, which also meant it became easier to mine at a larger scale. In the case of the cannabis industry, the mandate that each operator must have a secure and sound traceability system allows operators—who know how to access their data—virtually unlimited potential in performing higher level analysis. While RFM modelling is based on only three features, modern customer segmentation can involve several hundred or even several thousand features. As a result, the segments of the analysis become much finer, much richer to allow retailers to understand their customers at levels simply unattainable from RFM or demographic analysis<sup>17</sup>.

One of the more popular ways retailers have been able to acquire such specific data regarding their customers is through a loyalty program<sup>18</sup>. In a loyalty program, the customer benefits by receiving certain discounts, but the natural by-product<sup>19</sup> of the loyalty card is the data that the retailer can mine to better serve their customers and boost profits<sup>20</sup>. By using this data, retailers can create specific marketing campaigns, target certain customer segments with uniquely tailored discounts, or even invite old customers back into the store. This data allows for retailers to conduct ultra-specific marketing strategies that has transformed the way retailers compete in the age of Big Data.

In order to perform customer segmentation analysis at a high level, retailers have begun to incorporate aspects of machine learning into the analysis of their customers. More specifically, retailers are utilizing unsupervised machine learning tools such as clustering and dimensionality reduction to approach

---

<sup>16</sup>Marcus (1998, p. 495)

<sup>17</sup>Marcus (1998, p. 494)

<sup>18</sup>Cooil, Aksoy & Keiningham (2008, p. 13)

<sup>19</sup>The authors mention data being the natural by-product of specifically a user accessing a webpage, but the same principle holds.

<sup>20</sup>Su & Chen (2015, p. 2)

analysis in ways that cannot be matched without machine learning. Instead of focusing on only a few features or customers at a time, it is possible to write programs and implement algorithms that can take into account several more features or several more instances than traditional spreadsheets can hold or process. Because of this massive potential, retailers across all industries are attempting to leverage clustering algorithms such as K-Means or hierarchical clustering to more accurately and quickly segment their customers. The faster and better retailers are able to cluster their customers, the quicker they can market to them and thus acquire market share.

## 2.2 Challenges of Performing Analysis

The benefits of customer segmentation analysis are clear. By having a stronger understanding of their consumer base, retailers can properly allocate resources to collect and mine relevant information to boost profits. However, getting to the point of performing high-level customer segmentation analysis is more difficult than originally thought for many retailers. Many retailers may have the rights to the necessary data to perform the analysis, but do not have either the ability to access it in a user-friendly manner or have an employee that has the skillset to work with it. The lack of proper personnel or equipment to handle the necessary volume of data is perhaps the biggest hindrance to smaller firms being able to perform such analysis. The popularity of open-source programming software such as R or Python has certainly helped make this type of analysis more accessible, but it still would require retailers having someone on their team who can code in either of those languages. Additionally, some retailers are simply unaware of either the extent of their data collection or are not yet inspired to dig into it. Nevertheless, retailers that have not fully adopted customer segmentation analysis are likely not doing so simply because they cannot afford to spend the time, money, or labor to perform the analysis. Therefore, it is an aim of this paper to show that this rich analysis can be performed cheaply and efficiently.

However, there is a far subtler but still consequential reason why retailers do not implement customer segmentation analysis: it is too complicated to understand. When compared to traditional demographic segmentation or RFM analysis, high-level customer segmentation analysis requires far more precise knowledge of machine learning and the mathematics that describe how the algorithms work. In addition, traditional marketing analysts are not equipped with the math or programming skills necessary to successfully implement cus-

customer segmentation analysis with machine learning methods <sup>21</sup>; similarly, programmers and data analysts are not well-suited to handle marketing tasks. This poses another conundrum as it involves transforming a typical marketing assignment—segmenting customers based on purchasing behaviors—into a purely programming one, which means the marketing team does not have the skills to code it up themselves but the programming team does not have the marketing skills to interpret the results. Hence, there is a necessity for a hybrid role that involves knowledge of the business, programming, and marketing. In modern workspaces, this role is dubbed the data scientist or information specialist.

In sum, customer segmentation analysis is the process of trying to understand a consumer base by splitting it up into segments. While traditional analysts found some success with demographic or RFM analysis, these models simply do not have the technological capabilities to provide rich insight into more specific details regarding the customers. On the other hand, customer segmentation analysis that is combined with machine learning methods has the ability to transform the way a retailer thinks about their data. As such, retailers are trying to find cheap, easy ways to implement and communicate how clustering can be used to segment their customers.

Now that there has been plenty of introduction into customer segmentation analysis, it is time to take a look under the hood of some clustering algorithms before finally engaging in discussion of the analysis.

---

<sup>21</sup>Marcus (1998, p. 495)

### 3 Clustering Using Machine Learning Methods

While many applications of machine learning, such as regression and classification, focus on predicting the outcome or value of an instance, these applications do not attempt to understand similarities between instances, just the relationship between instances and their respective outputs. Thus, when it comes to searching for algorithms or methods that look for similarities between features of instances, the focus must turn from supervised machine learning to unsupervised machine learning.

Determining whether an algorithm is a part of supervised and unsupervised machine learning is contingent upon whether the instances used to train the model in the training data contain their target value. In all cases of supervised machine learning training, instances are paired with a target value, which could be a scalar or a vector depending on the context. In contrast, unsupervised machine learning deals with data that is not paired with a target value. To clearly spell out these differences — and also certain similarities — it may be best to examine them through an example.

For instance, consider a retail store owner who has a store that has been open for over a year and they are interested in examining their data to help boost understanding of their customers while also predicting how much they will spend next visit. To predict their next ticket, the owner takes their previous purchases and comes up with a way to guess, based on the previous tickets, the value of the next purchase. Since this example involves prediction and the outcomes of previous data and its outcomes (the tickets themselves), this is an example of supervised machine learning. To be more specific, since the owner is likely trying to predict a dollar amount the customer will spend, this type of algorithm is called regression.

On the other hand, to boost the understandings of their customers, the owner decides to look at some collected customer data and see if there are broader patterns or similarities between the customers. Since there is no clear outcome or target value associated with the data or the process, this is a type of unsupervised machine learning. More precisely, this exemplifies clustering.

In technical terms, clustering is an unsupervised machine learning technique that groups instances into clusters based on the similarities between instances. This just states that clustering is one way of viewing or evaluating data by looking at the natural groupings or segments that separate instances in the data. However, it is difficult to appreciate clustering without first fully understanding what it means for instances to be considered similar.

### 3.1 Similarity Measures

The success of a clustering algorithm rests upon the ability to choose the proper similarity measure before engaging in clustering. Choosing the best similarity measure, however, depends on an acute awareness of what similarity is and how it can be defined mathematically.

First and foremost, similarity in data science is a function of distance; the closer together two instances' values are, the more similar they will be. In certain contexts, defining distance between two instances is more obvious than others. If a data scientist were to cluster instances based solely on one numerical feature, then the clustering algorithm would take into account the differences between the instances and group them based on that. If the data scientist were to consider two features, the distance between features is a little more complicated. Instead of just the difference between the instances, we have the Euclidean Distance<sup>22</sup> between instances  $x$  and  $y$ :

$$distance(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (2D \text{ Distance Formula})$$

This formula comes from the Pythagorean Theorem and the fact that in Euclidean Geometry, the shortest distance between any two points is a straight line. The distance of the straight line in this case is calculated using the above formula. But, this is a good time to pause and evaluate certain understandings and motivations of what is going on here.

In order to find a way to compare how similar two instance are, it was first necessary to define the relationship between similarity and distance. In most contexts, the natural relationship to establish is an inverse relationship, which is what is used in this paper. The next necessity is to define the distance between two instances. With numerical data, such as the data in this project or in the previous examples, the natural distance measure to use is the standard Euclidean Distance Formula. The main reason why this is the natural measure for distance is that the data we are interested in clustering is numeric in nature. In other contexts such as Natural Language Processing, notions of similarity begin to diverge from the simple numerical notion presented here. But, the essential point is that finding the similarity between two instances involves at least two forethoughts; what is the relationship between distance and similarity, and how is distance defined in this context?

---

<sup>22</sup>It is worth noting here that the one-dimensional example is a unique case of this function with only one considered feature.



As alluded previously, this project makes use of Euclidean Distance as a way to define the distance between two instances. While the two examples above talk about distance on a one or two-dimensional level, the data in this project involves much more than two features, and so the intuition that guided the lower dimensional thinking needs to be expanded into higher dimensions. It turns out, when expanded into  $n$  features, the distance formula gets a more general look:

$$distance(x, y) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2} \quad (\text{n-D Distance Formula})$$

Many textbooks or academic papers may choose to refer to the distance formula in the context of clustering without the square root sign.

$$distance(x, y) = \sum_{i=0}^n (x_i - y_i)^2$$

This is a shorthand way of saving computational power in the actual clustering algorithm, but it ultimately is tied to the fact that when it becomes important to minimize the distance, finding the minimum of a squared distance or the square root of some squared distance yields the same minimum<sup>23</sup>.

Now that there has been a discussion on similarity, it is appropriate to begin to explore two different types of clustering algorithms. Both were used in the context of this project, which makes it crucial to compare the two algorithms in this paper because they reveal deeper insight into how different types of clustering can be used in differing contexts.

## 3.2 Centroid-based: K-Means

Although there are numerous types of clustering that each deserve their own mentions and explanations, this paper will focus on two types of clustering algorithms: centroid-based and hierarchical-based. Though, before beginning an ample discussion of centroid-based clustering, it is first necessary to understand what a centroid is and how they fit into clustering.

In the context of centroid-based clustering, a centroid is the center of a cluster of data. Although there are numerous ways to define the center of a cluster, the center in a k-means cluster is the arithmetic mean of each feature in the space in which the data exist. In other words, the centroid is the mean

---

<sup>23</sup>In situations where there are lots of data, taking the square root becomes a superfluous step in terms of both memory and time.

of the features of the instances that are assigned to that cluster<sup>24</sup>. However, it might not be immediately clear why centroids are necessary in the first place or how to initially define them. After all, there has not been any discussion on how exactly an algorithm would go about grouping data into clusters, let alone how centroids fit into that process.

To begin this discussion, it is appropriate to also begin with an example. Imagine that there is some 2D <sup>25</sup> data that, when plotted, looks like the following:

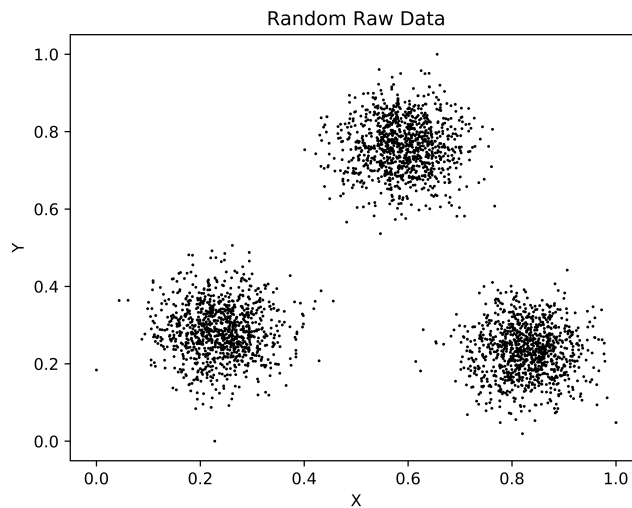


Figure 1: Random Raw Data

Immediately after looking at this figure, there are three things that are apparent. First, each data point exists in two dimensions:  $x$  and  $y$ . Although naming the axes  $x$  and  $y$  is convenient, it does not provide much insight into what the axes represent. So, if  $x$  and  $y$  are too simple, perhaps think of them as age and income or time between visits and average ticket. Regardless of what the axes' names are, the crucial point is that there are two dimensions.

Second, the data is scaled between 0 and 1 in both dimensions. Given this paper has yet to discuss the importance of scaled data in clustering, it might not make much sense why this is an important thing to note. Without revealing too much detail, the gist is that scaling keeps features that have large

<sup>24</sup>Rogers & Girolami (2016, p. 207)

<sup>25</sup>The math and intuitions of clustering extend into any number of dimensions, but it is often easier to distill the algorithm into two dimensions to investigate its underlying processes.

ranges from overwhelming data with smaller ranges. Furthermore, the 0-to-1 scale means that the maximum distance between any two points is  $\sqrt{M}$ , where  $M$  is the number of dimensions (in this case it is 2). To go along with this, the 0-to-1 scale also makes sure that none of the numbers, when squared, are bigger than 1; this is an often understated point in discussions of K-Means. When there are several hundred—or even thousand—numbers being summed and squared during the distance calculation, it is important to have smaller numbers because they will take up less memory in the long run. Thus, despite the 0-to-1 scale at first appearing meaningless, the scaling makes it easier to compute distance.

Lastly, to a human eye, there appear to be at least three distinct clusters. To some, this might be trivial to point out: by looking at the figure, it feels natural and also simple to place each data point into one of three clusters. In essence, this natural feeling is a reflection of the idea that humans are excellent at finding out patterns/commonalities<sup>26</sup> between instances under two conditions: when there are not that many instances and when there are not that many features. In the figure, there are two features and although there are several thousand instances, plotting them all at once makes it easier to see the differences between each datum. Because of the low number of features and the ability to see all the data clearly, the human mind has little difficulty dividing up the data into clusters. However, teaching a computer to perform the same task is slightly more difficult. For all the incredible tasks that a CPU can perform, it cannot visualize the data and divide it into nice groups like a human can. Therefore, it is reasonable to wonder how a computer would go about the task of clustering.

In centroid-based clustering, the most popular algorithm is k-means. There are two important parts to the name:  $k$  and means. Here,  $k$  refers to the number of centroids (clusters) the algorithm will generate and “means” refers to what the centroids are: arithmetic means of the data<sup>27</sup>. Roughly the k-means algorithm can be broken up into four sections, each with their own important attributes.

To start, while it is clear what a centroid is, it is unclear how it fits into the algorithm at the beginning. First, if a centroid is supposed to be the arithmetic mean of the points that belong to it, how is it possible to use them initially? In other words, how does one know where to put the centroids? In short, the smartest and most common decision to make is to randomly place

---

<sup>26</sup>Mattson (2014, p. 265)

<sup>27</sup>Rogers & Girolami (2016, p. 207)

the centroids throughout the dataset<sup>28 29</sup>. Here, it might help to think of centroids as points in the same space that the data belong. In the raw data shown in figure 1, centroids would appear as a random point between 0 and 1 for both its x and y component. In relation to this, it is also reasonable to wonder how many centroids to randomly place throughout the dataset. For the sake of simplicity, let's initialize three centroids and place them randomly<sup>30</sup> throughout the dataset.

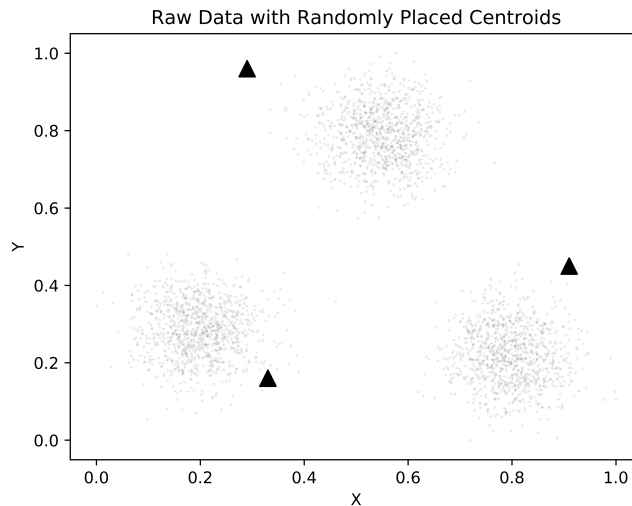


Figure 2: Raw Data with Centroids

Now that the centroids exist in the space, it is nearly time to begin clustering. Before beginning the main chunk of the k-means algorithm, it is necessary to assign each point to one —and only one— of the centroids. To assign a point to a centroid, one must first find the distance from each point to each centroid. The data point, thus, will be assigned to the centroid that is closest to it or, equivalently, the one to which it is the most similar. Figure 3 shows an example of taking one point and computing the distance (shown in red) between itself and each of the three centroids. From the figure, it becomes easy to see that the randomly selected point should be assigned to the left-most

<sup>28</sup>Rogers & Girolami (2016, p. 207)

<sup>29</sup>Wagstaff, Cardie, Rogers & Schrödl (2001, p. 578)

<sup>30</sup>In truth, the centroids here were not truly randomly placed and the decision to place three was equally not random. They were placed so that they were spaced out enough to provide a clear example. Though, since it is often impossible to have a good guess of where to place the centroids, it is smart to truly randomly place them.

centroid, since it is the closest centroid to the point. This process of assigning points to the closest centroid repeats for all remaining points in the dataset.

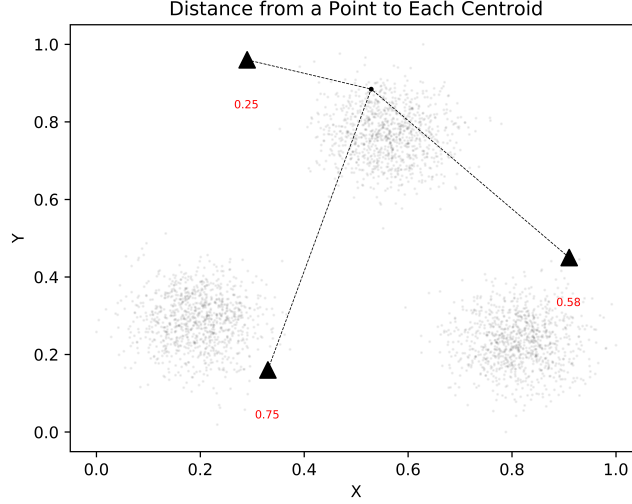


Figure 3: Distance from a Random Point to Each Centroid

Once each point is assigned to a centroid, it is time to update the position of each centroid. In k-means, recall that the centroid is the arithmetic mean of the data that belong to that centroid. So, in two dimensions, this idea can mathematically expressed as:

$$centroid_i = \frac{1}{n} * (\sum_{j=1}^n X_{(j,1)}, \sum_{j=1}^n X_{(j,2)}) \quad (2D \text{ Centroid Update})$$

Here, each instance resides in  $X$  and instance  $X_j$  is assigned to centroid  $i$ . Any instance not assigned to centroid  $i$  does not affect the reassignment of the centroid. Furthermore, the terms  $X_{(j,1)}$  and  $X_{(j,2)}$  indicate the value of the first and second features of instance  $X_j$  respectively. Lastly, the  $\frac{1}{n}$  is the way this calculation becomes an average, since  $n$  represents the number of instance belonging to centroid  $i$ .

However, it is also generally useful to understand how similar concepts can be applied outside of two dimensions. When expanded into higher spaces, the update formula changes to (in  $k$  dimensions):

$$centroid_i = \frac{1}{n} * (\sum_{j=1}^n X_{(j,1)}, \sum_{j=1}^n X_{(j,2)}, \dots, \sum_{j=1}^n X_{(j,k)}) \quad (k-D \text{ Centroid Update})$$

Figure 4 graphically shows the placement of the new centroids after updating their positions. When comparing this figure to figure 2, it is evident that each of the centroids moved toward the direction of the points closest to them.

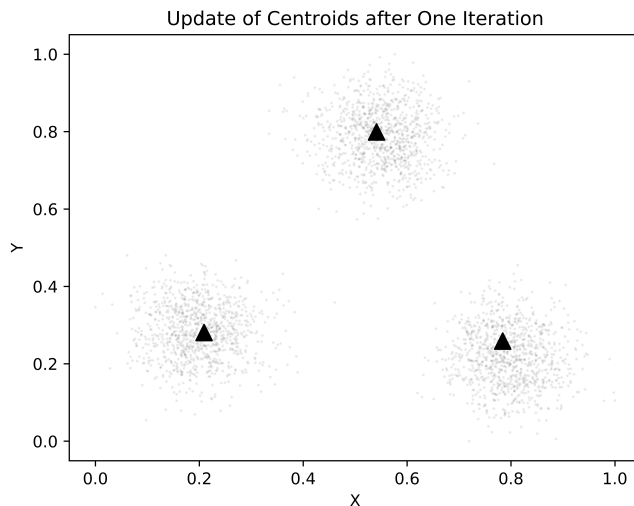


Figure 4: Update of Centroids after One Iteration

The update of centroids, now, is not too difficult to explain or implement, but when should the algorithm stop updating centroids? In practice, the k-means algorithm stops when either the centroids remain unchanged from the previous iteration or, equivalently, the labelling of each point to a centroid remains unchanged from the previous iteration<sup>31</sup><sup>32</sup>; this is called convergence. Since the data in this example are particularly well-suited for clustering, it should not be a surprise that the first iteration of k-means yields centroids that are very much close to the ideal centers of each of the clusters. Similarly, the algorithm here, as shown in figure 5 actually converges only after two iterations, a very fast convergence<sup>33</sup>.

<sup>31</sup>Wagstaff, Cardie, Rogers & Schrödl (2001, p. 578)

<sup>32</sup>Usually, these two conditions will happen at the same time. It is worth mentioning both though because it may be easier to check one condition or another based on the context or setup of the problem.

<sup>33</sup>k-means is intended to be an algorithm that converges quickly, but it would very rarely converge in two iterations unless the data already had a nice, cluster-like structure to it.

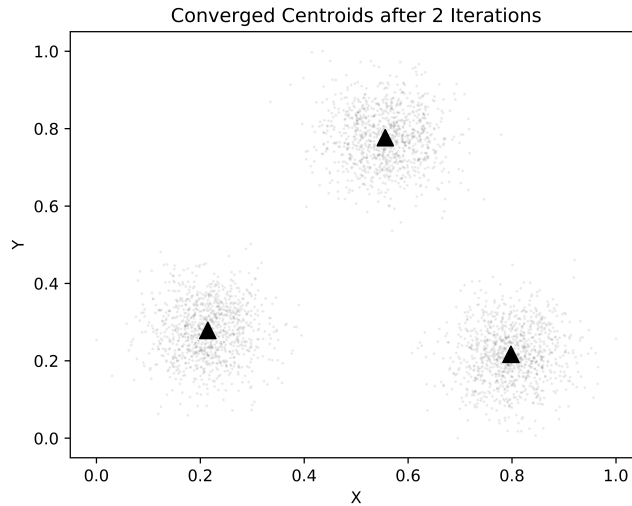


Figure 5: Converged k-means Algorithm

In this particularly simple example, the centroids converged rather quickly mainly due to the fact that the centroids were already close to nicely shaped clusters. But, consider a case where the centroids are not nicely placed, perhaps closer to each other or closer toward the middle of the data. In this case, the centroids do not converge anywhere near as quickly, and they may converge in different spots than the simple example. Because of this, it is common for data scientists to run the K-Means clustering with several different random initial assignments and take the one that has the smallest inertia, which is the sum of the square distance between each point and the centroid. Inertia will be discussed more specifically in section 5 in the context of choosing the optimal  $k$ —or the number of centroids—to cluster with.

Before continuing, it is worth the time to quickly summarize k-means and centroid-based clustering. The k-means algorithm works by taking  $k$  centroids and randomly placing them across the dataset, ideally so that they are evenly spaced out. Each datum then gets assigned to the centroid it is closest to, which is the centroid with the smallest Euclidean Distance between it and the datum. Once all the data have been assigned, the centroids update by becoming the mean of all the data assigned to it. When the centroids stop moving or the assignments stop changing, the algorithm stops. To get close to the optimal solution for a particular  $k$ , it is recommended to rerun the algorithm with different initial centroid assignments.

In sum, centroid-based clustering is one of the most common ways to cluster data with machine learning methods, but it is not flawless. For example, the

number of centroids,  $k$ , has to be chosen beforehand, which makes it more difficult to find the optimal  $k$  to cluster with. Furthermore,  $k$ -means operates under the assumption that the data will have nice “centers”<sup>34</sup> to their segments that will allow the centroids to converge, which is often not the case with real-world data<sup>35</sup>. This also implies that  $k$ -means is sensitive to outlier data that can cause centroids to converge far from the optimal spot. So, for as powerful and influential as  $k$ -means is, there are clear reasons to explore other strategies that do not suffer from the same weaknesses. Thus, it is necessary to dive into an alternate form of clustering known as hierarchical clustering.

### 3.3 Hierarchical-based: Agglomerative

Centroid-based clustering is extremely popular but often improperly applied to real-world datasets. In addition to being susceptible to outliers, it is also frustrating for analysts to determine the proper number of centroids,  $k$ , to specify beforehand. To circumvent these shortcomings in practice, it is common to explore a different kind of clustering algorithm.

While centroid-based clustering is intuitive and easy to implement, hierarchical clustering is comparable in its implementation but does not suffer from the drawbacks that centroid-based clustering does. In short, hierarchical clustering is a type of clustering based on either a top-bottom or a bottom-top approach. More specifically, a bottom-up approach—where each datum starts as one cluster until they all merge into one giant cluster—is known as agglomerative clustering<sup>36</sup> whereas the converse is known as divisive clustering<sup>37</sup>. Because agglomerative clustering is more intuitive to explain than divisive, this paper will make use of it.

Unlike the centroid-based clustering, agglomerative clustering begins with each point acting as its own cluster with the end goal of each cluster eventually merging with other clusters. Figure 6 displays a random initialization of some data where each point will be its own cluster.

---

<sup>34</sup>While there are several rigorous ways to define the significance of a “nice center”, the simplest to explain is that the features in the data should be normally distributed such as in the provided example data.

<sup>35</sup>Su & Chen (2015, p. 2)

<sup>36</sup>Ward Jr. (1963, p. 238)

<sup>37</sup>Roux (2018, p. 347)



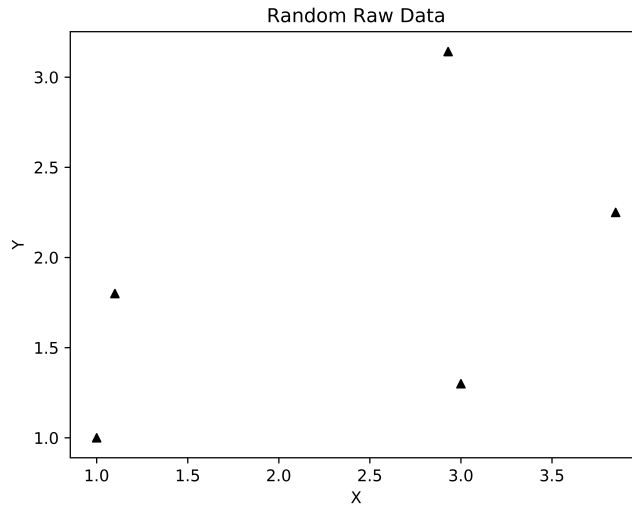


Figure 6: Random Raw Data for Agglomerative Clustering

In any clustering project, it is important to develop some intuitions about the structure of the data prior to engaging in clustering. When looking at the data, the two points in the bottom left corner should stand out due to their relative closeness. In a similar fashion to centroid-based clustering, it is important to explicitly define what “close” means in this context. When each point is its own cluster, it is common to use the Euclidean Distance formula if the features are all numeric such as here. So, it is natural that the first step of the agglomerative clustering algorithm is to find the two closest points: in this case, it is the two leftmost points. In order for the computer to find the two closest points, it is necessary to compute the distance from each point to every other point. This is most easily accomplished by creating a distance matrix where position  $(i, j)$  represents the distance between points  $i$  and  $j$ . The distance matrix is important because it holds the key to the order of grouping for the data. Lastly, while the order of grouping for the rightmost points might not be immediately clear, it should be simple to see that the first grouping will occur with the leftmost points and move on from there. Figure 7, below, visualizes a sample distance calculation for one random point.

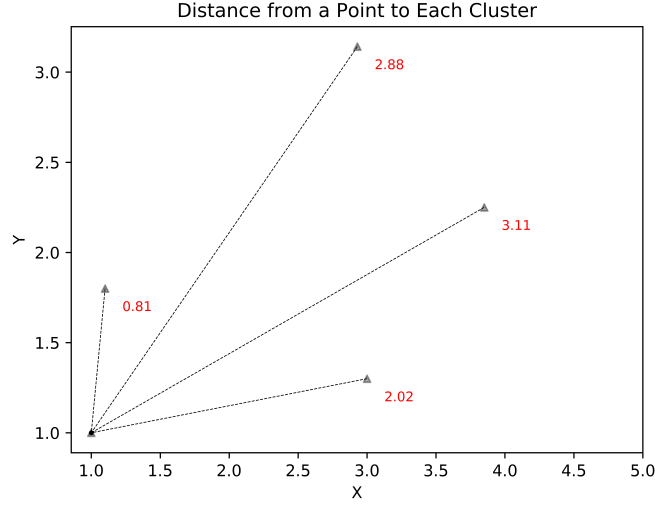


Figure 7: Sample Distance Calculation from One Point to Each Cluster

Once the shortest distance is found between two clusters, the two clusters merge into one cluster. In this example, there were five initial clusters but after one iteration, there are four remaining clusters. In general, the pattern of the number of clusters starts at  $n$ , goes to  $n - 1$ ,  $n - 2$ ,  $\dots$ , 2, then finally 1. With one merged cluster, it now becomes a little trickier to define how this merged cluster should act in the specified distance formula. In essence, the three most common methods of defining this interaction are to:

1. Take the shortest possible distance between one of the points in the merged cluster and the desired point. This is the defining characteristic of the single linkage method.<sup>38</sup>
2. Take the largest possible distance between one of the points in the merged cluster and the desired point. Conversely, this is often referred to as the complete linkage method.<sup>39</sup>
3. Take the distance from the desired point to the center of the merged cluster. Simply put, this is often called the average linkage method.<sup>40</sup>

Without loss of generality, the rest of the example will utilize the average linkage method. This is perhaps best shown in figure 8, where the square in the middle of the merged cluster represents the centroid.

<sup>38</sup>Rajaraman & Ullman (2011, p. 242)

<sup>39</sup>Tan, Steinbach, Karpatne & Kumar (2018, p. 555)

<sup>40</sup>Tan, Steinbach, Karpatne & Kumar (2018, p. 555)

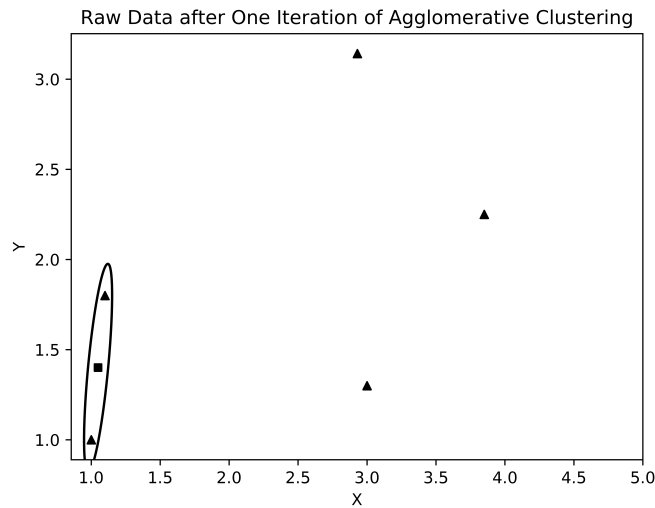


Figure 8: Example of a Data Update in Agglomerative Clustering

In future iterations of the algorithm, each point will compare itself to the centroid of the merged cluster. As the algorithm progresses, the grouping of clusters and subclusters naturally forms a tree-like structure. This structure, known as a dendrogram, displays crucial information regarding the way the algorithm clustered the data. For this particular example, the results are summarized below in figure 9.

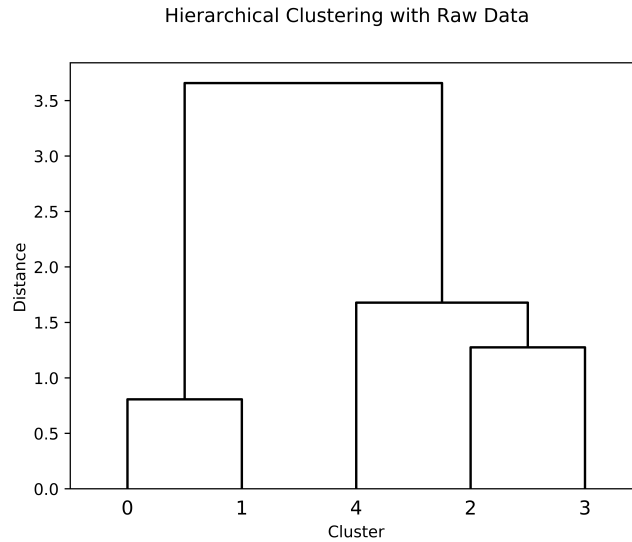


Figure 9: Dendrogram of Clustered Random Raw Data

When analyzing a dendrogram, it is important to examine the locations of the connections and how their distance from the closest connection, which appear as the flat lines on the figure. For example, points 0 and 1 — the leftmost and two closest points — join first and thus have the connection closest to the bottom. The shorter the distance between the connections, the more closely related the two clusters are <sup>41</sup>. Once all the clusters finish agglomerating, each of the connections and their relationship to one another is beautifully clear.

To recap, hierarchical clustering is a type of clustering that involves establishing a hierarchy in terms of how similar two clusters are. In agglomerative clustering, each datum starts as its own individual cluster and proceeds to join with the most similar cluster. With numeric data and working with Euclidean Space, the Euclidean Distance Formula is the most common distance/dissimilarity metric. However, since the number of clusters updates each iteration of the algorithm, it is important to keep track of the distance between each cluster to every other cluster via a distance matrix. Once there is only one cluster remaining, the algorithm converges and the results are commonly visualized with a dendrogram.

.....  
Clearly, the previous discussions of clustering and customer segmentation

---

<sup>41</sup>Tan, Steinbach, Karpatne & Kumar (2018, p. 558)

analysis show that these two topics are important for retailers to understand. While deserving of their own research and recognition, these topics are far better understood when they work in conjunction with one another. Clustering can transform a retailer's customer segmentation analysis into a formidable research tool by providing richer analysis and utilizing more data. In fact, the harmony between customer segmentation and clustering is precisely the relationship that overarches the motivations and implications of writing this paper. Without further ado, it is finally time to begin discussing the setup and results of the clustering experiment performed for 4Front.

## 4 Preparing the Data

The digressions of clustering and customer segmentation analysis were important, but it is now time to think back to the previously stated business problem and the associated data. Although several variables from each data table were listed, not all the variables could be used in the analysis as is. Certain variables, such as the ID columns, provide necessary information to corroborate data and keep accurate calculations between instances, but are not necessarily features that merit analysis<sup>42</sup>. On a similar note, features, such as the time of a specific transaction and the value of its ticket, contain essential information for mining, but need to be transformed into a more usable format. In particular, these transaction-based variables need to be converted into customer-based variables. However, variables such as the product category are on an item-based level, which require a separate transformation of their own. Nonetheless, the salient point is that it is necessary to consider the raw data, examine its format and original features, and transform them into a workable format for the task at hand.

### 4.1 Feature Engineering

The process of creating or extracting features from raw data is commonly referred to as feature engineering. Often, it is the first and most important step of data preprocessing because it establishes the features that the model will consider when clustering. Essentially, feature engineering involves inspecting and manipulating the raw data to somehow extract features that are worthwhile for analysis. Because the concept of a “worthwhile” feature is subjective, the data scientist must place the task’s mission and constraints at the forefront of their decision-making process with regard to engineering features. In this project specifically, one of the main goals is to obtain a better understanding of 4Front’s customers based on their purchasing patterns. So, the features that will appear on a customer-based level, describe purchasing patterns, and extract the most information from the raw data will be optimal features for the project. After poring through the datatables, there were 11 unique features that were engineered that are summarized as follows:

1. Age<sup>43</sup>: the age of the customer as of July 10, 2019

---

<sup>42</sup>Since each ID is randomly formed at the time of creation, there is no pattern or relationship between one customer’s ID and another.

<sup>43</sup>There was much debate about including age, a demographic variable, but leaving out other demographic variables such as race or sex. The decision to omit other demographic

2. Visits: the total number of visits the customer has made to the store since inception
3. Total Spent: the total amount spent across all visits
4. Average Ticket: total spent / visits
5. Average Time between Visits: the average number of days between visits. By default, it is -1 if the customer has visited less than twice
6. Flower: the proportion of purchased items that are categorized as flower
7. Vape: the proportion of purchased items that are categorized as vape, which includes live resin catridges as well as the standard distillate cartridges
8. Concentrate: the proportion of purchased items that are categorized as concentrate. Included in this is kief, shatter, wax, batter, and other dabbable forms of cannabis<sup>44</sup>
9. Preroll: the proportion of purchased items that are categorized as a preroll. This includes individually packaged prerolls as well as those sold in packs of two or more.
10. Edible: the proportion of purchased items that are categorized as edibles. Chocolates, drinks, teas, gummies, and mints are examples of what is considered an edible.
11. Topical/Other: the proportion of purchased items that are categorized as either topicals or anything not in the above categories. These two are smashed together into one feature because topicals, like items not in the other categories, are sold far less than other items.

While most customer segmentation analyses focus on the RFM model, the features here provide profound insights into the purchasing behavior of the customer. Rather than solely analyzing them based off the amount spent and

---

variables was based on their ultra-discrete nature as well as the fact that a considerable chunk of all customers did not list their sex or race, but all customers had included their age. This likely is a direct result of the mandatory customer registration process on their first visit, which requires new customers to report their age.

<sup>44</sup>If these words seem fake or unrecognizable, I recommend perusing some popular online resources such as Leafly. They provide comprehensive information on all types of cannabis.

visits, the cannabis-related data drives the segmentation to focus on cannabis-specific purchasing behavior. This is crucial to recognize because there is not significant research of cannabis retail data <sup>45</sup>. Thus, there is motivation to perform analysis in a way that is directly applicable to cannabis data.

Lastly, computing or finding the necessary data for analysis was more difficult than hypothesized, particularly finding the cannabis-related purchase behavior data. In the 4Front database, each product has a specific product category ranging from 1-20. However, there is insufficient documentation on what each of the product categories refer to. As a result, it was necessary to take a sample from each product category and manually classify them into the broader categories (Flower, Edible, etc.) used in the clustering. Moreover, if this analysis is to be repeated across stores, the product categorization would have to be completely redone<sup>46</sup>. This presents a problem for cannabis retail data scientists that will only likely be addressed if similar struggles are publicly shared.

## 4.2 Criteria for Clustering

To expedite the clustering process, the new customer data needed to undergo minor data preprocessing. In this step, certain customers were pruned from the dataset if they did not meet certain self-imposed constraints. At the particular dispensary studied, there were 15,489 unique customers as of July 10, 2019 that had spent a total of \$3,743,454. However, only 4,975<sup>47</sup> (32.12%) of the customers were able to be clustered in the dataset. These 4,975 were chosen for meeting the following criteria:

- The customer checked “Yes” to allowing their data be used for internal and marketing purposes
- Their birthday and other necessary data had no malformed values. The birthday deserves special mention because some birthday inputs had only two digit years or months bigger than 12, which made it hard to absolutely determine their age. Less than 100 of the several thousand customers failed this criteria

---

<sup>45</sup>Morrison, Gruenewald, Freisthler, Ponicki & Remer (2014, p. 508)

<sup>46</sup>Despite consulting numerous employees of 4Front with direct retail knowledge, there is still no consensus on where the categories originated. It is currently hypothesized that each state mandates a particular categorization system and forces POS systems to comply.

<sup>47</sup>For those interested in the application of the Pareto Principle, these 32.12% of customers accounted for \$2,936,161 (78.43%) in total sales



- The customer must have visited at least three times. This is to ensure that there is ample data collection and that time-based features, such as average time between visits, can be meaningful.

While there is only around a third of the original dataset remaining, the data is now dense enough to allow for customer preferences to truly appear. In turn, the results of the clustering will be much richer than having include the whole dataset, provided that there is as little Survivorship Bias<sup>48</sup> as possible.

### 4.3 Scaling and Reformatting Data

As mentioned previously, it is often a good idea to have data scaled between 0 and 1 before engaging in clustering. This is true because it prevents the distance formula from accumulating computationally taxing sums, since each term of the sum is between 0 and 1. Scaling data between 0 and 1 is relatively straightforward:

$$feature_i = \frac{feature_i - \min(feature)}{\max(feature) - \min(feature)} \quad (\text{MinMax Scale Formula})$$

Where *feature* is the feature that is becoming scaled. It is important to scale features rather than instances in this context because instances are the focus of comparison, not the features; in other words, we are clustering instances, not features.

While it would have been preferred that all data would work well with a simple MinMax scaling, one drawback of MinMax scaling is that it is very susceptible to outliers. Certain features such as total spent and average time between visits vary so widely across customers that a MinMax scale would not be adequate in the sense that it would not mitigate the variability in the data. Thus, it is often common to apply a log transform (or some other transform such as a power) to the data, then MinMax scale the transformed data instead. This achieves the ultimate purpose of scaling— to get the data between 0 and 1— while circumventing the issue of outliers. To display it clearly, here are the transformations applied to the following features:

- Age - MinMax scale

---

<sup>48</sup>Survivorship Bias arises during any data pruning, where the pruned dataset contains fundamentally different patterns/behaviors than the original. Here, this bias is mitigated because the research question is related to repeat customers, not necessarily all customers as a whole

- Visits -  $\text{Log}_2$  transform, then MinMax scale
- Total Spent -  $\text{Log}_{10}$  transform, then MinMax scale
- Average Ticket -  $\text{Log}_{10}$  transform, then MinMax scale
- Average Time between Visits -  $\text{Log}_{10}$  transform, then MinMax scale
- All other features are already within a 0-1 range

Once the data passed the criteria for clustering and was scaled/reformatted, the data was then prepared for clustering. The following section describes these results and the implications of them.

## 5 Performing Analysis and Results

After the data was formatted in an appropriate way, it was time to begin clustering the data. While the clustering algorithms were implemented using `sklearn` — a popular, open-source Python data science library — there was significant coding needed to not only get to the point of clustering, but also recording the results in a reasonable manner. Hence, it is only proper to first provide an overview of the programming needed to create the workflow.

### 5.1 Brief Overview of Code

The entirety of the code written for this project was in Python. The following Python packages played pivotal roles in the execution and development of this project:

- `pandas`, `numpy`, `sklearn.preprocessing`, `os`, `datetime`, and `time` were all used for data collection, handling, and manipulation
- `seaborn`, `matplotlib.pyplot`, and `scipy.cluster.hierarchy` were used to create visualizations of data
- `sklearn.cluster` and `sklearn.decomposition` were used for clustering the data or decomposing it into three dimensions for plotting <sup>49</sup>

In general, the dataflow consists of five separate steps. First, the raw data collected from the database is cleaned for malformed values, voided tickets, and items that were not sold <sup>50</sup>. If necessary, this data can be saved and stored for future access. Next, the remaining data is turned into customer-based data. Each unique customer is initialized with their purchase data encapsulated by the features used for clustering. Additionally, customers that do not meet the criteria for clustering are pruned from the dataset, leaving only customer data that is able to be clustered. Once the customer data are formed, the data are then scaled and reformatted via the reformatting procedure laid out in section 4.3.

The reformatted data are now ready to be clustered. Consequently, the next step is the clustering of the data with both K-Means as well as Agglomerative. Since the clustering is performed with high-level, open-source packages

---

<sup>49</sup>This is referring to Principal Component Analysis (PCA), a method used to visualize higher-dimensional data.

<sup>50</sup>In some states, each dispensary is required to offer the customer educational materials on cannabis. Most customers decline the materials, but as proof, an item in each ticket is “\* Declined Educational Materials”. Thus, these need to be removed.

such as sklearn, the clustering is extremely fast, regardless of the number of clusters chosen. The results of the clustering are saved into a variety of locations based on the format of the results; results that involve labelling the raw data are moved into a separate location than the data that describes the structure of the clusters. Furthermore, some data on the runtime or other meta results of clustering were collected. Altogether, the dataflow, when done in its fullest form, takes around eight minutes to complete, which is far from optimal.

## 5.2 Clustering Results

### 5.2.1 K-Means

The first clustering performed in the dataflow was K-Means. Since K-Means requires a prespecified  $k$  to cluster, K-Means was run with  $k$ s from 1-25 to ensure an ample range for sufficient clustering to occur. Tied to this, each iteration of clustering was run with randomly initialized centroids 100 times, with the best <sup>51</sup> clustering chosen from each one. The results of the clustering are summarized in figure 10.

---

<sup>51</sup>The best clustering is the one that minimizes the inertia, or the total sum of squares from each point to its centroid

Results of k-Means Algorithm on [REDACTED] Dispensary Customer Data with Various ks

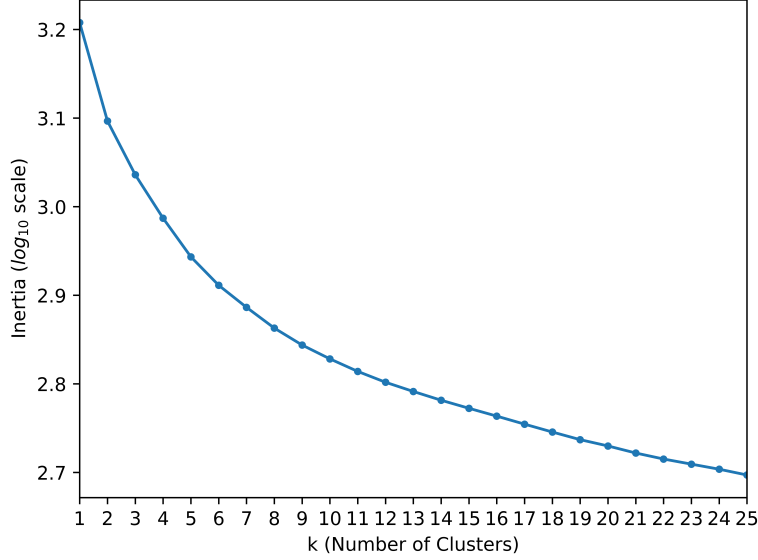


Figure 10: Inertia for Several Ks on 4Front (dispensary name redacted) Dispensary Data.

While typical examples of K-Means yield nice elbow curves, this clustering does not. There is no clear, discernible elbow point on the curve in figure 10 that indicates an optimal k. Some argument can be made that one occurs around 6 or 7, but this is not conspicuous. Perhaps the optimal k is beyond the range, but that would likely involve results that are significantly overfitted or not actionable. Since the result of the optimal k is less obvious than desired, it was decided to explore clustering with five clusters. Five was chosen because it was small enough to be actionable but large enough to provide specific breakdowns within the clustering. As a result, the statistics and means of each feature for the five clusters are reported in table 1.

Table 1 reveals the natural segmentations of the customers. First off, the clusters are not as balanced as hoped, but are still close to within 10% of the expected (20%). The most popular clusters, clusters two and four, are the heaviest vape and flower consumers, respectively. In other words, these are the variables that most clearly delineate these clusters. While cluster two has much higher tickets and total spent, cluster four consumes far more prerolls than cluster two. Due to the low number of visits and high time between visits, cluster four are likely customers that do not shop consistently, but perhaps bulk

Table 1: Results of K-Means Clustering with  $k = 5$

Feature/Cluster	1	2	3	4	5
Count	588(11.82%)	1,163(23.38%)	755(15.18%)	1,505(30.25%)	964(19.38%)
Age	32.7	35.7	35.9	36.0	37.1
Visits	7.45	6.92	32.56	5.21	6.23
Total Spent	\$453.86	\$542.37	\$1,733.12	\$270.75	\$334.57
Ticket	\$63.94	\$78.85	\$59.20	\$53.27	\$55.53
Time between Visits	46.59	45.84	12.49	50.75	44.97
Flower	20.57%	19.39%	52.76%	70.63%	25.46%
Vape	12.63%	56.54%	12.09%	5.99%	13.07%
Preroll	8.15%	8.60%	13.09%	12.05%	23.13%
Edible	5.69%	5.75%	9.00%	4.49%	27.35%
Topical/Other	0.28%	0.53%	0.52%	0.43%	1.53%
Concentrate	48.82%	4.57%	6.89%	3.18%	4.82%

up on cheap flower deals. On the other hand, customers within cluster two are spending more money more frequently, which makes sense given a 1-gram vape cartridge can be near \$40 while one gram of flower is near \$10 on average. These segments are worth highlighting not only because they are the largest in terms of size, but they are the base consumers of two of the most popular types of products at a dispensary: vapes and raw flower.

Despite consuming less vape and flower than clusters two and four, perhaps cluster three is the most intriguing cluster. Each of the other four clusters averages between 44 and 50 days between visits, but consumers in cluster three are visiting more than three times as often as the other clusters. And while their tickets are not significantly higher, their visits and thus total spent are the highest of any cluster. To make this cluster even more peculiar, this cluster also is not the highest in any of the cannabis-specific features. This suggests that they dabble in each of the product types offered at the dispensary. In a sense, this makes them the “connoseuir” cluster, which means that they are also likely the most loyal and educated customers. Their connoseuir character is also vastly similar to cluster five, which has more even balanced cannabis-specific features but is more defined by their edible consumption in addition to their low tickets.

Last but not least, concentrate consumption and age best separates cluster one from the rest of the clusters in the data. The low age and high concen-

trate consumption may suggest that cluster one represents the younger, newer cannabis consumers that lean towards dabbing concentrates rather than purchasing flower or vapes such as in clusters two through five. In a similar sense to cluster two, cluster one has a high average ticket because the average price for gram of concentrate is much higher than flower. Interestingly enough, this is also the smallest cluster by several hundred customers, so it may not be as influential to the total customer breakdown as the other clusters.

In sum, centroid-based K-Means yielded five clusters that were mostly differentiated by their purchasing behavior and cannabis-specific behavior rather than their demographics. While clusters two and four are the largest clusters in terms of size, cluster three is the cluster with the highest total spend and the most visits. Lastly, cluster one, the youngest cluster, consumes mostly concentrates, cluster five consumes mostly edibles, clusters three and four consume mostly flower and prerolls, and cluster two consumes mostly vapes.

### **5.2.2 Agglomerative**

Once K-Means clustering finished, agglomerative clustering was performed immediately after. Recall that agglomerative clustering produces a dendrogram, which is the main tool that the data scientist will use to decide how to determine the optimal number of clusters. Figure 11 shows the dendrogram formed prior to clustering. After looking at the dendrogram, the decision was made to cut at a point that created six customer segments, shown by the figure-wide horizontal line. While there is a considerable argument for cutting with five clusters, six here was chosen because it was far enough down the tree to produce discernible clusters but not high up enough to make the clusters too generalized.

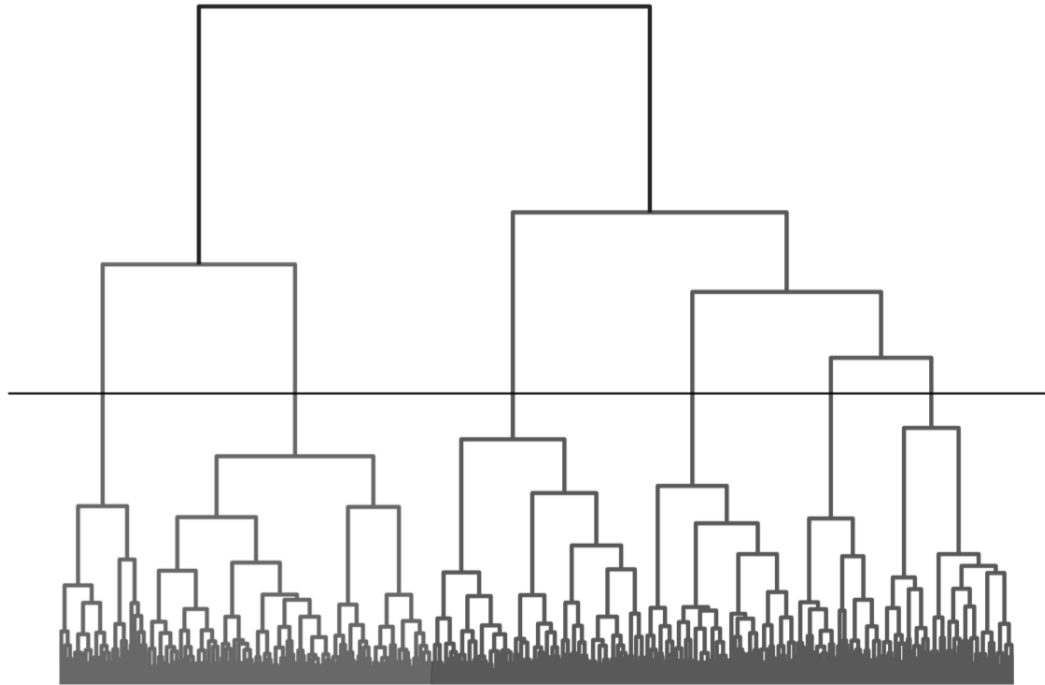


Figure 11: Dendrogram for 4Front Dispensary Data

Unlike the inertia plot created during K-Means, the dendrogram is more comprehensible and easier to explain to non-data scientists, which is one of the reasons why the 4Front management preferred investigating the results of agglomerative clustering. Furthermore, when looking at this particular dendrogram, the right side seems to be significantly more broken down than the left. While it is not known where each cluster lands on the dendrogram, it is nice to visually see the relationship between each cluster, especially considering that this is not as easily accomplishable with K-Means. In any case, the characteristics of the clustered data are presented in table 2.

When first looking at the table, it is noticeable that the segmentations are not evenly populated, which leads to a structure, such as this, where more than 50% of the customers fall into two segments (2 and 3) and another two segments (4 and 6) only account for 18% of the customers. This imbalance



Table 2: Results of Agglomerative Clustering with 6 Clusters

Feature/Cluster	1	2	3	4	5	6
Count	691(13.89%)	1,491(29.97%)	1,111(22.33%)	456(9.17%)	787(15.82%)	439(8.82%)
Age	46.1	35.1	33.4	35.6	31.9	34.8
Visits	7.08	5.97	7.45	42.04	6.16	10.86
Total Spent	\$447.01	\$317.39	\$584.73	\$2,185.88	\$397.69	\$443.44
Ticket	\$64.59	\$54.78	\$77.04	\$56.71	\$65.86	\$41.25
Time between Visits	39.48	43.86	50.60	10.31	47.93	42.27
Flower	27.77%	70.39%	22.24%	53.63%	21.71%	33.99%
Vape	18.69%	6.94%	53.44%	10.37%	17.18%	6.98%
Preroll	10.66%	10.54%	9.68%	9.34%	9.30%	44.86%
Edible	32.21%	5.45%	6.55%	8.50%	7.20%	6.25%
Topical/Other	2.10%	0.45%	0.45%	0.38%	0.37%	0.49%
Concentrate	4.50%	2.82%	3.00%	11.37%	39.57%	3.86%

is not necessarily detrimental to the analysis because there are still relevant distinctions between each of the clusters. For example, clusters two and three are the most popular clusters with two key differences: cluster two, the largest cluster, are thrifty consumers (\$54.78 average ticket) that predominately purchase flower, which accounts for over 70% of their purchases on average; on the other hand, cluster three consumers mostly purchase vapes, consequently leading to the highest average ticket of all clusters at \$77.04. In addition, the thrifty nature of cluster two is evident by having the lowest number of average visits (5.97) and in turn the lowest average total spent. It is also peculiar that clusters two and three have two of the three highest values for any cannabis-specific variable, which must indicate that flower and vape purchases comprise a significant portion of a significant number of consumers at the dispensary.

While flower and vape consumption best delineated clusters two and three from the rest, cluster one’s separation is primarily due to age. While the other clusters are around 32 to 35 in age, the average age for cluster one is 46.1. The severity of the gap in age between cluster one and the rest is surprising because it seems to indicate that older consumers have a unique purchasing behavior from younger consumers; this is also evidenced by cluster one having the largest edible and topical/other consumption compared to younger clusters consuming more flower or concentrate. Yet, it is noteworthy that cluster one has the most balanced purchase profile, with its highest category peaking at 32.11%. Lastly, cluster one is a relatively normal size, unlike clusters two, three, four, or six, if one considers that the average cluster size should be

around  $\frac{1}{6}$  or 16%.

In an interesting manner, the other normally sized cluster, cluster five, is dichotomous to cluster one in numerous ways. To begin, cluster five is the youngest cluster whereas cluster one is the oldest. Even though the clusters have mostly similar average tickets, flower consumption, and vape consumption, they are vastly different on edible and concentrate consumption. The age split between cluster one and five is revealing, but perhaps the cannabis-behavior split between the groups is more insightful. This supplies evidence to ideas and marketing that associate younger cannabis users with concentrates and older users with topicals and edibles.

In terms of spending features, cluster four is the most intriguing cluster of the six. The average visits and total spent are vastly different from the rest of the clusters in a way that is almost incomparable, which is certainly remarkable. Even though it is nearly the smallest cluster, it accounts for the highest total contributed revenue<sup>52</sup> in part because of the low time between visits (around 10 days) and high average visits. Like cluster two, cluster five consumes mostly flower, but otherwise they are not a leader in any of the cannabis-specific categories. This suggests that consumers in this cluster are familiar with different forms of cannabis, which may indicate that these consumers have holistic knowledge and broad experience with cannabis.

Finally, cluster six, the smallest cluster, is primarily defined by preroll usage. Because prerolls are usually some of the cheapest products that dispensaries sell, it is not entirely surprising that cluster six has the lowest average ticket (\$41.25) but also a relatively high number of visits. This may point towards an idea that prepackaged products such as vapes, edibles, and prerolls require more frequent visits rather than non-prepackaged products such as raw flower or concentrates. Like most of the other clusters, cluster six has around 40 days between visits and a fairly average age. Nonetheless, the main characteristics of cluster six are its preroll consumption and its relatively high average number of visits.

In summary, tables 1 and 2 independently convey the results of the different clusterings and are clearly different in numerous ways. Most conspicuously, there are differing number of clusters between the tables, which makes one-to-one comparison impossible. However, that is not to say that the tables are not similar in any way. In fact, the two tables more or less express the same information. They both communicate the importance of flower and vape consumption, as they are the defining characteristics of the largest clusters in

---

<sup>52</sup>In this context, the total revenue contributed is the average total spent multiplied by the size of the cluster.

both tables. Interestingly enough, cluster two in table 1 and cluster three in table 2 nearly have identical sizes and also both were clustered primarily on vape consumption, just like clusters four and two in tables 1 and 2, respectively. Likewise, there is a unique cluster in both tables that is devoted to the ultra-high level consumers. This is positive because it means that these distinctions are pertinent, so much so that multiple clustering algorithms recognize them as defining traits.

### 5.3 Managerial Implications of Results

Tables 1 and 2 describe important patterns and behaviors of the dispensary’s consumers, but there are crucial implications that accompany these results, particularly with regard to the structure of the clusters. To begin, analysis of both algorithms indicates that the optimal number of clusters to choose is somewhere between five and six. With only five or six clusters, it is straightforward to separate out the clusters to uncover patterns. However, a lower number of clusters might be easier to separate, but the clusters will be far less informative and too general to make accurate predictions. This is often a problem with traditional customer segmentation analysis. Since there is motivation to keep analysis simple and not expand features unless necessary, traditional customer segmentation analysis often leads to oversimplification of the clusters and thus complicates managerial action. Regardless, the decision to cluster with five or six segments is present throughout customer segmentation analysis research<sup>53 54</sup>. In one sense, this implies that clustering cannabis retail data, even with cannabis-specific variables, may not be different from clustering other retail data. In turn, applying methods performed with other types of retail data to cannabis retail data is not only applicable, but perhaps even recommended as both the size and complexity of the data evolve.

Besides the difference in the number of clusters, the inherent purchasing behavior remains very much intact between both clusters. For example, clusters four and two in tables 1 and 2 are almost identically the same size and have nearly equal values for their features; this is also the case with cluster two in table 1 and cluster three in table 2. The parallelism between the two tables should not be entirely surprising, since they are clustering with the same data. Yet, the prevalence of the specific purchasing profiles in both tables reveals how easily detectable these profiles are from the given data. When cluster profiles are consistent across numerous different clustering algorithms,

---

<sup>53</sup>Shepitsen, Gemmell, Mobasher & Burke (2008, p. 263)

<sup>54</sup>Chen, Sain & Guo (2012, p. 203)

it indicates that the profiles are more than just quirks in the data: they are real signals that emerge from the noise. For managers, this translates to confidence that the profiles obtained from clustering algorithms are meaningful and actionable.

When the profiles obtained from customer segmentation analysis are meaningful, they can be converted into real marketing campaigns that target customers based on their purchase profiles, rather than traditional demographic profiles. One of the common problems with traditional demographic segmentation is that it is too simple to describe the convoluted purchasing nature of a retailer's customers. When clustering data with variables that are domain-specific, there is a clear intention to uncover domain-specific patterns that are unattainable with RFM or demographic segmentation. These uncovered patterns are invaluable to a retail company because they inherently communicate a retailer's business strategy. If a retailer is able to find meaningful patterns within a clustering, the patterns can be used to form actionable customer profiles. These actionable profiles make it easy for retailers to trust the results of the clustering, which turns into deployment into the quotidian business strategy. A strong trust between retailers and data scientists is integral to the success of any data-based retail project. In essence, it cannot be understated how using domain-specific data in clustering can form segmentations that are actionable and thus profitable.

While most customer segmentation analyses focused on traditional RFM analysis, it has been noted throughout this paper that there are stronger, more insightful ways to engineer relevant features for segmentation analysis. However, it is not useful to simply add features that are irrelevant for the sake of adding them. In order to come up with informative features, it is often necessary to have domain-specific knowledge to engineer proper features; the success of any machine learning project depends on the features available to it. So, deriving features is not a trivial task, but it is imperative that it is done correctly. The combination of cannabis-specific domain knowledge and practical skills of data science helped create features in this project that elegantly define each cluster. Yet, these features were not created *ex nihilo*: they came from taking a deeper look into the raw data that is already collected within the database. Although many retailers and operators within the cannabis industry view the strict traceability as a burdensome necessity, it offers a goldmine of opportunity in terms of data. By encouraging deeper looks into the raw transactional data, retailers can develop more actionable and profound profiles of their customers and products

## 6 Future Work and Conclusion

### 6.1 Possible Research Avenues or Expansions

While there has been plenty of time devoted to discussions of customer segmentation analysis, machine learning, and the results of clustering with cannabis retail data, is time to revisit the end of the first section of the paper included a list of four goals that were important to accomplish for the paper to fully cover its scope. Goals one, two, and four were achieved in the first five sections, but goal three requires its own special attention. More specifically, there needs to be discussion into not only ways to improve the current project, but also ways to expand upon it or use its ideas or findings in other contexts.

As it stands, there are at least four visible improvements that could be made to the current project. First, perhaps most obviously, there should be more clustering algorithms used to fully understand the range of customer profiles and also the number of segments within the customer data. Although there was plenty of information gleaned from just two clusterings, different cluster algorithms can communicate additional findings or handle different sets of constraints. As mentioned previously, K-Means is the most popular clustering algorithm but it suffers from the key drawbacks of requiring the number of centroids to be established a priori in addition to requiring variables to be numerical in nature. Although hierarchical clustering fixes this problem, it is not as scalable as K-Means and also assumes an inherent hierarchical structure of the data. Furthermore, neither of the algorithms provide a probability of an instance belonging to a particular group; they both simply classify the instances into clusters. An additional clustering algorithm that can provide probabilities of belonging to a cluster is called Gaussian Mixture Models. With a probability of cluster assignment, retailers can begin to look at clustering as less rigid of a process. Ultimately, this offers a more flexible approach to marketing and profiling rather than strict clustering algorithms such as the ones in this paper.

Indeed, exploring other clustering algorithms is worthwhile, but another important potential improvement is to enhance the data to make more precise clusters. While the project aimed to accomplish goals of traditional customer segmentation analysis, there was no variable to account for the recency of the customer, mostly because there was confusion over defining it. Since the dispensary had been opened for less than two years at the time of the analysis, it proved difficult coming up with an objective way to measure the recency of a customer. The original thought was to, like the other variables, scale the number of days since last visit between 0 and 1, but this creates an in-

comprehensible structure: a higher value would mean less recent. There were discussions of turning the recency into a categorical variable (have they visited within the last two months), but this would prove to be more trouble than it is worth because of the struggles—and sensitivities—that the relevant clustering algorithms have with categorical data<sup>55</sup>. So, the motivation to add a recency feature is justifiable, but the implementation of it is far more difficult than envisioned.

Another relevant implementation to the current project would be to add additional measures of cluster stability and validity. Clustering, in its very essence, is a way to explore data; naturally, clustering projects tend to focus more on investigating the patterns that arise in the data rather than evaluating rigorous loss or benefit metrics such as in supervised machine learning. As clustering research continues to expand, many data scientists have proposed a variety of measures or tools to address common concerns of clustering. Some common ways to evaluate clustering include computing the silhouette coefficient, creating a proximity matrix, turning the clustering results into a decision tree and computing the entropy/purity, and calculating the inertia or SSE of the model<sup>56</sup>. While these measures do not tell the whole story, they can illuminate the strengths or limitations of the present clustering architecture. In the end, this leads to a holistic comprehension of the data and results.

On a lesser note, the data collection process can be improved, not necessarily for results but for efficiency. The original code, while passable, struggled to prune the raw data in a timely fashion. After inspecting the bottlenecks of the code, it became clear that one of the issues involves updating a dataframe each iteration of cleaning rather than updating all at once; instead of updating the dataframe, in its entirety, one time, the data flow now updates the entire dataframe several thousand times, which is slower than it should be. Fixing the runtime of a workable project should be the final update before publication or deployment, so there is not urgent motivation to correct these bottlenecks as of now.

Once the proper improvements are made, there are numerous ways to combine the ideas and results from the clustering with other data projects within 4Front. Although this analysis focused on the results of only one dispensary, 4Front operates numerous dispensaries across the country. Fine-tuning the

---

<sup>55</sup>A fair criticism of this point is that one-hot encoding the categorical variable would not make it “categorical.” However, K-Means works best with continuous, dense variables. One-hot encoding does turn a categorical variable into a numerical one, but it does not make it continuous. In fact, the discrete nature of one-hot encoded features can complicate the results of K-Means or hierarchical clustering.

<sup>56</sup>Tan, Steinbach, Karpatne & Kumar (2018, p. 587-595)

analysis to one store is not optimal. Rather, the dataflow should be as generalizable as possible. On a practical level, this means discovering how to collect the same features from other POS systems, which may require a far more involved process than the one used in this project. But once consistency in the dataflow is achieved, clustering customers in different markets may generate advanced understanding of the economics of the area, or even the more general customer segmentation in the cannabis industry. In short, expanding this analysis into each of the available markets gives 4Front a competitive advantage that can be leveraged to beat out competition.

Exploring the customer data at one point in time can provide enormous insight, but it is also possible to explore the evolutionary clustering of the customers. One method is to see how the size and optimal number of clusters evolve with time. In particular, studying the evolution of the optimal number of clusters may hint at a deeper understanding of how customers naturally segment within a retail setting. These results can then be compared to evolutionary cluster studies in other industries to answer a longstanding question within the cannabis industry: do cannabis customers act differently than customer in other retail industries? Simply taking a look at the results of evolutionary clustering is not only worthwhile for 4Front, it could be groundbreaking for the industry.

Lastly, clustering is not specific to customer segmentation analysis: it can be used in any segmentation project. With that said, clustering other relevant data within the retail database is also an expansion to the current project. Although the feature engineering process and dataflow would be altered, the general path of analysis would remain the same; convert the raw data into a usable format (e.g. day-based or product-based), scale and reformat the data appropriately, and implement the algorithms. Depending on the basis of the data, the results of the clustering uncover patterns not just within customers but also certain days of operation or even groups of products. As a side effect, this also promotes further discussion of machine learning in everyday retail analysis and therefore makes the concepts and practices of data science more easily understood in the business setting.

## 6.2 Conclusion

For the most part, the cannabis industry is in its nascent stages. The intense federal criminalization of cannabis for years totally hampered professional research into all facets of cannabis, from cultivation to retail to consumption. As a result, dispensaries are learning how to navigate not just a thicket of regulations and other constraints, but also an unclear road of consumer behavior.

Conducting direct research with consumers and products is not possible, so retailers must look inward to uncover the behaviors of their customers. Despite many retailers outside of cannabis have had tremendous success with traditional customer segmentation analysis, the supply of skilled analysts willing and capable to serve the cannabis industry is far smaller. To compound this, there is plenty of data in the cannabis industry— due to the enforcement of a traceability system— but few ways to access it. Although dashboards and elaborate interfaces have eased the responsibility of finding patterns or commonalities in retail data, none of them provides statistics or data that is rich enough to make advanced insights such as customer segmentations. As a result, it is necessary to bring in tools that are specifically built for situations such as this: machine learning.

With ample data, cannabis-specific domain knowledge, and a background in machine learning, developing a set of scripts to cluster the raw data was possible. After engineering relevant features and reformatting the data, it was possible to perform customer segmentation analysis with two different clustering algorithms: K-Means and Agglomerative. Even though the algorithms used different numbers of clusters in their clusterings, they essentially convey the same three pieces of information. First, flower and vape consumption were the defining characteristics of the largest clusters, which hints at the importance of these two products to a dispensary’s success. Second, both algorithms generated a cluster of ultra-frequent consumers, with average visits and total spent significantly higher than the rest of the clusters. Lastly, the tables also show that older consumers tend to enjoy edibles and topicals more than other consumers; on the flip side, younger consumers tend to enjoy vapes and concentrates more.

Regardless of the information provided, the results provide actionable ways for retailers to employ a marketing campaign or similar segmentation for their consumers. Despite the usefulness of the analysis as-is, there are numerous routes for improvement and growth. While there was motivation to keep the number of features low, adding a separate feature to account for the recency of the consumer would provide clearer details on whether certain purchase profiles are more common now than in the store’s past. On a similar note, finding ways to cluster a customer quicker (such as in one or two visits rather than three) could generate insights into not only the evolutionary aspect of the clustering but potentially also the leakage of customers. Finally, attempting the same analysis with numerous other clustering algorithms such as Gaussian Mixture Models or deep learning would bring about insight into the stability of cluster formation.

.....



This project could not have been completed without the gracious wisdom of my thesis advisor, Dr. Richard Thompson, as well as the level-headedness of my boss, Joe Feltham, from 4Front Ventures. Finally, I would also like to extend a neverending gratitude towards my parents, Randall and Gina, and my partner, Rebekah: your love, motivation, and passion for my success is inspiring.

## References

- Bhatnagar, Amit; Ghose, S. (2004), ‘A latent class segmentation analysis of e-shoppers’, *Journal of Business Research* **57**, 758–767.
- Chen, D., Sain, S. L. & Guo, K. (2012), ‘Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining’, *Journal of Database Marketing & Customer Strategy Management* **19**(3), 197–208.  
**URL:** <https://doi.org/10.1057/dbm.2012.17>
- Cooil, B., Aksoy, L. & Keiningham, T. L. (2008), ‘Approaches to customer segmentation’, *Journal of Relationship Marketing* **6**(3-4), 9–39.
- Marcus, C. (1998), ‘A practical yet meaningful approach to customer segmentation approach to customer segmentation’, *Journal of Consumer Marketing* **15**, 494–504.
- Mattson, M. P. (2014), ‘Superior pattern processing is the essence of the evolved human brain’, *Frontiers in Neuroscience* **8**, 265.
- Morrison, C., Gruenewald, P. J., Freisthler, B., Ponicki, W. R. & Remer, L. G. (2014), ‘The economic geography of medical cannabis dispensaries in california’, *International Journal of Drug Policy* **25**(3), 508 – 515.  
**URL:** <http://www.sciencedirect.com/science/article/pii/S0955395913002387>
- Rajaraman, A. & Ullman, J. D. (2011), *Mining of Massive Datasets*, Cambridge University Press, New York, NY, USA.
- Rogers, S. & Girolami, M. (2016), *A First Course in Machine Learning, Second Edition*, Chapman & Hall/CRC.
- Roux, M. (2018), ‘A comparative study of divisive and agglomerative hierarchical clustering algorithms’, *Journal of Classification* **35**(2), 345–366.  
**URL:** <https://doi.org/10.1007/s00357-018-9259-9>
- Shepitsen, A., Gemmell, J., Mobasher, B. & Burke, R. (2008), Personalized recommendation in social tagging systems using hierarchical clustering, in ‘Proceedings of the 2008 ACM Conference on Recommender Systems’, Rec-Sys ’08, ACM, New York, NY, USA, pp. 259–266.  
**URL:** <http://doi.acm.org/10.1145/1454008.1454048>

- Su, Q. & Chen, L. (2015), ‘A method for discovering clusters of e-commerce interest patterns using click-stream data’, *Electronic Commerce Research and Applications* **14**(1), 1 – 13.  
**URL:** <http://www.sciencedirect.com/science/article/pii/S1567422314000726>
- Tan, P.-N., Steinbach, M., Karpatne, A. & Kumar, V. (2018), *Introduction to Data Mining (2nd Edition)*, 2nd edn, Pearson.
- Wagstaff, K., Cardie, C., Rogers, S. & Schrödl, S. (2001), Constrained k-means clustering with background knowledge, *in* ‘Proceedings of the Eighteenth International Conference on Machine Learning’, pp. 577–584.
- Ward Jr., J. H. (1963), ‘Hierarchical grouping to optimize an objective function’, *Journal of the American Statistical Association* **58**(301), 236–244.
- Zhang, G. (2007), Customer segmentation based on survival character, *in* ‘2007 International Conference on Wireless Communications, Networking and Mobile Computing’, pp. 3391–3396.